cadernos BAD

7ª conferência luso-brasileira acessoaberto Instituto Politécnico de Viseu - Portugal

Gestão de dados: Sobreposições ou convergências entre infraestruturas?

Pedro Moura Ferreira
Instituto de Ciências Sociais, Universidade de Lisboa, APIS
pedro.ferreira@ics.ulisboa.pt

Bárbara Rodrigues
Instituto de Ciências Sociais, Universidade de Lisboa, APIS
barbara.rodrigues@ics.ulisboa.pt

Resumo

No domínio da gestão de dados, o debate sobre a estratégia mais adequada para a curadoria e preservação dos dados de investigação em Portugal tem ganho uma pertinência crescente. A experiência acumulada nesta área aponta para as vantagens de um modelo em rede que articule *soluções de âmbito genérico*, lideradas pelas estruturas de coordenação central, e *soluções de âmbito disciplinar*, dinamizadas pelas infraestruturas de investigação de acordo com as necessidades e desenvolvimentos das diferentes áreas científicas. Nesta comunicação, a atividade dos arquivos de dados em ciências sociais é usada como exemplo para demonstrar as vantagens de conciliar estas duas abordagens.

Palavras-chave: Curadoria de Dados; Dados de Investigação; Gestão de Dados; Infraestruturas de Investigação; Preservação Digital.

Data management: overlapping or convergence between infrastructures?

Abstract

In the field of data management, the debate concerning a suitable strategy for the curation and preservation of research data in Portugal has become increasingly pertinent. Previous experience points to the advantages of a network-based model that combines a general approach, headed by the central coordination structures, with a disciplinary approach, driven by the research infrastructures in line with the

needs and changes of the different scientific areas. In this paper, the activities of social sciences data archives will be used to demonstrate the advantages of combining these two approaches.

Keywords: Data Curation, Research Data, Data Management, Research Infrastructures, Digital Preservation.

Introdução

Em Portugal, a curadoria dos dados de investigação constitui um dos campos insuficientemente desenvolvidos do sistema científico e tecnológico. Com efeito, são ainda limitadas as iniciativas neste âmbito, nomeadamente ao nível de infraestruturas que suportem a preservação e a disseminação dos dados. Por sua vez, as iniciativas existentes são bastante diversas, apresentando ora um âmbito *nacional*, se considerarmos o RCAAP; *disciplinar*, no caso das infraestruturas de investigação; e, ainda, *institucional*, no caso de ações como as desenvolvidas pelos Serviços de Documentação da Universidade do Minho e Universidade do Porto.

Tendo em vista responder às atuais e, sobretudo, futuras necessidades de depósito, arquivo, preservação e disseminação de dados de investigação, a articulação entre os diversos atores constitui um ponto que merece ser discutido e aprofundado, tanto mais quando um conjunto diversificado de fatores, como os requisitos de acesso aberto, a exigência de transparência dos processos de investigação, a reutilização dados e os receios de perdas de informação, têm vindo a contribuir para uma nova atitude dos investigadores face aos dados de investigação, ao mesmo tempo que lhes é exigido mais competências em termos de gestão de dados (Corti et al, 2014).[i] Esta mudança no perfil das responsabilidades e competências dos investigadores tem de ser acompanhada pelo desenvolvimento de infraestruturas de dados de investigação, porque, conforme a experiência de outros comprova, a existência dessas mesmas infraestruturas estimula e aprofunda o princípio da reutilização dos dados, contribuindo para maximizar o benefício público da investigação.

Objetivo

Neste sentido, esta comunicação tem por propósito identificar e explorar as possíveis convergências, complementaridades e eventuais sobreposições entre os repositórios institucionais de dados de investigação e o papel que as infraestruturas de investigação, designadamente as que integram o no Roteiro Nacional de Infraestruturas de Investigação de Interesse Estratégico 2014-2020 da Fundação para a Ciência e Tecnologia, poderão desempenhar ou vir a desempenhar relativamente à gestão dos dados de investigação. O objetivo é esboçar possíveis articulações entre os diferentes tipos de infraestruturas, chamando a atenção para a necessidade de uma moldura institucional coerente e, tanto quanto possível, sob uma orientação unificada e integrada, decorrente de uma política científica relativa aos dados de investigação.

Metodologia

Do ponto de vista metodológico, a relação entre os repositórios de dados de investigação e as infraestruturas de investigação é analisada e discutida no âmbito das ciências sociais, em particular a partir do caso da infraestrutura *Production and Archive of Social Science Data* (PASSDA), que reúne o Arquivo Português de Informação Social e o Programa das Atitudes Sociais e Políticas dos Portugueses. A análise realizada teve naturalmente em consideração a experiência e a documentação produzida quer no âmbito das infraestruturas, quer no âmbito dos repositórios dos dados, em especial a que foi realizada pelo RCAAP, sem ignorar as contribuições de importantes eventos como as Jornadas da Fundação para a Computação Científica. Considerando, no entanto, o propósito de clarificação do papel dos repositórios disciplinares associados às infraestruturas de investigação torna-se imprescindível recorrer à experiência internacional, especialmente relevante no que se refere aos arquivos de dados de ciências sociais reunidos no consórcio CESSDA e ao qual se espera que Portugal venha a aderir.

Resultados e discussão

O Repositório Científico de Acesso Aberto de Portugal constitui um projeto-chave do sistema científico e tecnológico nacional e o principal dinamizador do movimento de acesso aberto em Portugal. A missão do projeto é «promover, apoiar e facilitar a adoção do acesso aberto ao conhecimento científico em Portugal e armazenar, disponibilizar e preservar a produção científica» (RCAAP, 2016).

Inicialmente comprometido com a disponibilização de literatura científica, o projeto alargou a sua atividade aos dados de investigação em 2010. A importância do RCAAP como um protagonista da gestão de dados em Portugal foi revigorada em eventos recentes como as Jornadas para a Computação Científica Nacional de 2016 e a Conferência de Dados de Investigação e Ciência Aberta - Rumo a uma Estratégia Nacional no mesmo ano.

Apesar do forte movimento político atual em direção à ciência aberta, é ainda pouco clara a distribuição de papéis e responsabilidades ao nível da gestão de dados em Portugal. Num estudo elaborado no âmbito do projeto RCAAP (Rodrigues, 2010), considerava-se que pelo menos a curadoria dos dados deveria ficar a cargo de profissionais, ao invés dos investigadores. O perfil destes profissionais era considerado ainda pouco definido. Em termos de infraestruturas de suporte, os repositórios institucionais das universidades eram vistos como uma das possíveis respostas para o alojamento de dados, disponibilizando, por exemplo, acesso aos dados produzidos pela «pequena ciência» (Rodrigues, 2010, p.17), constituindo deste modo um *upgrade* dos serviços já existentes. Ou seja, estes repositórios estariam mais vocacionados para dados de alcance mais limitado junto da respetiva comunidade científica, enquanto os dados mais relevantes ou mesmo de maior qualidade estariam a cargo de centros de dados.

Independentemente das soluções referidas, o estudo desaconselhava fortemente a criação de mega soluções para o arquivo de dados, propondo antes uma solução de

equilíbrio entre um *modelo de base genérica* e um *modelo de base disciplinar* (Rodrigues, 2010). Segundo este documento, «a curadoria dos dados científicos, para ser verdadeiramente efetiva e sustentável, exige a participação de todas as partes (os investigadores, as instituições onde trabalham e os organismos de financiamento) envolvidas na produção dos dados e no processo de investigação» (Rodrigues et al., 2010, p.42).

No contexto de um modelo de base disciplinar, os arquivos europeus de dados em ciências sociais representam uma das soluções com mais tradição na gestão de dados de investigação. Os primeiros arquivos foram criados na década de 60 e 70 do século passado – primeiro na Alemanha, seguida dos EUA, Reino Unido e Noruega – pela mão de investigadores empenhados na criação de estruturas que disponibilizassem recursos para a análise comparada e para a análise secundária de dados em ciências sociais. Foi assim que surgiu também, em 1976, o atual *Consortium of European Social Science Data Archive* (CESSDA), (Kaase, 2013), que em 2016 assumirá o estatuto de infraestrutura europeia de investigação (ERIC). O CESSDA tem 24 arquivos associados, dos quais 15 são membros (CESSDA, 2016). A existência desta rede internacional tem permitido uma ação concertada na gestão de dados de investigação, através da partilha de boas práticas, normas e conhecimentos técnicos, mas também na diminuição das barreiras entre países no acesso a recursos de investigação de qualidade.

Orientados para a o apoio à investigação, os arquivos oferecem um conjunto diversificado de serviços além, naturalmente, da função de pesquisa e de disponibilização dos dados, como o apoio à elaboração de planos de gestão de dados, consentimentos informados ou aconselhamento ético, banco de questões, abarcando por isso diferentes fases do ciclo de vida dos dados, desde a sua criação até à sua reutilização. Em muitos casos, os arquivos trabalham em estreita colaboração com as infraestruturas de investigação, como será o caso do Arquivo Português de Informação Social já que faz parte da infraestrutura *Production and Archive of Social Science Data Archive* (PASSDA).

As infraestruturas de investigação em ciências sociais têm ganho um protagonismo crescente nos sistemas científicos e tecnológicos, podendo ser definidas como "instituições duráveis, ferramentas técnicas e plataformas, e/ou serviços implementados para apoiar e promover a investigação enquanto recurso público para a comunidade de ciências sociais" (Renschler, Kleiner e Wernli, 2013, p.1). O investimento em infraestruturas constitui uma matéria consensual ao nível europeu patente na criação de instrumentos como o *European Strategy Forum on Research Infrastrucutres Roadmap* (ESFRI) e o *European Research Infratrucutre Consortium* (ERIC). Tais instrumentos conferem às infraestruturas as condições necessárias para a estabilidade na prestação de serviços ao longo do tempo. A produção e a localização de recursos para a promoção da investigação de excelência, a inovação das práticas e a forte internacionalização das suas redes, constituem alguma das características destas infraestruturas.

Os arquivos asseguram a continuidade do acesso aos dados produzidos pelos projetos de investigação através de um conjunto de atividades de curadoria, preservação e disseminação. Ocupam-se assim de coleções, que incluem, para além das bases de dados, documentação como questionários e relatórios metodológicos.

Na medida em que o objetivo do arquivo de dados é permitir o uso futuro dos mesmos e não somente a sua preservação ou replicação, os arquivos desempenham um conjunto de tarefas diversificadas em comparação com o repositório institucional. Entre estas tarefas, destaca-se a documentação detalhada da informação depositada pelos investigadores com vista a facilitar a sua reutilização.

> «Ao pensar na integração de bases de dados abertas, é inevitável abordar a questão dos *standards* e dos metadados. Apenas com esta preocupação com a disponibilização de dados será possível criar uma rede de fontes que permitam a interoperabilidade de dados e que não constitua um "caos informacional", no qual seja impossível aos investigadores guiarem-se nas buscas pela informação pretendida» (Cardoso, 2012, p. 32).

Os arquivos europeus de dados em ciências sociais baseiam a sua documentação no *Data Documentation Initiative* (DDI). O DDI é um *standard* internacional para a descrição estruturada de dados em ciências sociais, criado em 1995. De acordo com Wackerow e Vardigan (2013), existe um conjunto diversificado de motivos para a utilização de metadados estruturados como seja (1) a preparação de dados para a sua reutilização por terceiros para análise secundária; (2) uma descrição compreensiva da coleção de dados; (3) relacionar estudos longitudinais; (4) o suporte à preservação, já que se baseia na linguagem XML. O DDI é também interoperável, não proprietário e garante a preservação a longo prazo pelo que está mais imune a perdas de informação.

A documentação dos dados encontra-se fortemente relacionada com a função de preservação digital dos arquivos de dados em ciências sociais, permitindo-lhes garantir o acesso às suas coleções no longo prazo.[ii] Porém, além desta documentação, existem ainda outras estratégias de preservação da informação que incluem, por exemplo, a migração de formatos, *backups*, verificação de autenticidade de ficheiros (ex.: *fixity checks*), melhorias nas infraestruturas técnicas, gestão de segurança e avaliação de riscos.

Neste contexto, a utilização de modelos de referência, como o *Open Archival Information System* (OAIS), é comum entre os arquivos do CESSDA. O OAIS é um modelo de referência, passível de ser utilizado por qualquer organização que tenha dados a seu cargo, que define as relações entre as principais atividades e funções relacionadas com a preservação digital – *ingest function, archival storage, data management function, administration function, access function* (Lavoie, 2000).

Outro aspeto fundamental no desenvolvimento dos arquivos tem sido o fortalecimento da confiança junto dos investigadores, financiadores e utilizadores de dados, através da certificação dos arquivos. Atualmente, seis dos 15 arquivos do consórcio CESSDA estão certificados pelo *Data Seal of Approval* (DSA). Este «selo» tem por objetivo garantir a

segurança dos dados e uma gestão de dados fiável e de qualidade, baseando-se numa autoavaliação, posteriormente revista por pares (DSA, 2016).

Conclusões: convergências entre infraestruturas

A breve incursão realizada aos arquivos europeus permitiu chamar a atenção para algumas das especificidades disciplinares relacionadas com a gestão de dados de investigação. No caso das ciências sociais, um modelo disciplinar de gestão de dados apresenta vantagens do ponto de vista da (1) proximidade com os investigadores e respetivas infraestruturas de investigação, (2) adequação dos *standards* aos dados disponibilizados, (3) documentação compreensiva, (4) conhecimento de boas práticas, (5) coordenação internacional das inovações e das boas práticas. E, ainda, no acompanhamento de necessidades relacionadas com a investigação que se estendem para além dos dados em acesso aberto; por exemplo a disponibilização de facilidades para análise de dados sensíveis recolhidos por organizações como os institutos nacionais de estatísticas.

Importa, no entanto, referir que uma estratégia nacional – como aquela que está em curso para a Ciência Aberta em Portugal – deve ter também em consideração uma ação concertada entre as diferentes áreas disciplinares. A formação de competências na gestão de dados, a articulação com a política científica nacional e a garantia de persistência dos objetos através da atribuição de DOI aos dados de investigação são exemplos de assuntos de interesse para qualquer infraestrutura de investigação. O ponto, porém, mais crítico para qualquer infraestruturas de investigação, relaciona-se com a preservação digital dos seus objetos no longo prazo. Os dados de investigação e a sua efetiva reutilização requerem que se dê particular atenção à preservação digital de modo a neutralizar os riscos de perda de informação que decorrem da obsolescência tecnológica ou da gestão inadequada dos dados. Só garantindo o acesso continuado aos dados é possível criar condições que satisfaçam a reutilização dos mesmos, contribuindo ao mesmo tempo para uma maior eficiência do sistema de Ciência & Tecnologia, no sentido em que se evita a duplicação de estudos e se promove a análise secundária.

Tendo em conta a necessidade de conciliar interesses comuns e disciplinares, faz sentido referir propostas avançadas no passado no âmbito da atividade do RCAAP e que poderão ser reconsideradas num futuro próximo, se esse for o entendimento prevalecente, agora considerando explicitamente os dados de investigação. Estas propostas sugeriam a criação de um grupo para a preservação e curadoria digital que, entre outros, concretizasse «um projeto piloto, no domínio da preservação digital, com a participação de vários repositórios portugueses, com o recurso a arquitetura(s) que possa(m) dotar os repositórios participantes de ferramentas abrangentes em termos de preservação digital»; avaliasse a «exequibilidade e os termos de uma possível cooperação entre o projeto RCAAP e/ou os repositórios individualmente com o RODA – Repositório de Objetos Digitais Autêntico»; desenvolvesse ou disseminasse «documentos de divulgação, formação e suporte, como

Briefing papers, modelos de políticas e procedimentos, boas práticas e casos exemplares de preservação digital» (Ferreira et al., 2012, p. 45-46).

A concretização desta iniciativa pressupõe não só aprofundar um diálogo já iniciado entre os vários protagonistas mas também a elaboração de orientações claras no que respeita à construção das infraestruturas e à gestão dos dados, as quais reclamam e exigem definições mais claras e previsíveis por parte das instâncias responsáveis pela condução da política científica.

Referências bibliográficas

CARDOSO, Gustavo; JACOBETTY, Pedro; DUARTE; Alexandra (2012) - *Para Uma Ciência Aberta*. 1ª ed. Lisboa : Mundos Sociais. 126 p. ISBN 978-989-8536-07-5.

CESSDA. *National Data Services.* [Em linha]. Bergen: CESSDA AS. [Consult. Set. 2016]. Disponível na Internet: <URL:http://cessda.net/National-Data-Services>.

CORTI, Louise [et al.] (2014) - *Managing and Sharing Research Data: A Guide to Good Practice*. 1ª ed. Londres : Sage. 222 p. ISBN 978-1-4462-6725-7.

DSA. Information. *About*. [Em linha]. [Consult. Out. 2016]. Disponível na Internet: <URL:http://www.datasealofapproval.org/en/information/about/>.

FCT (2014) - *Portuguese Roadmap for Research Infrastructures*. [Em linha]. [Consult. Abr. 2016]. Disponível na Internet: <URL:https://www.fct.pt/apoios/equipamento/roteiro/2013/docs/Portuguese_Roadmap_of_Research_Infrastructures.pdf>.

FERREIRA, Miguel; SARAIVA, Ricardo; RODRIGUES, Eloy (2012) - *Estado da Arte em Preservação Digital*. RCAAP. [Em linha.] [Consult. Set. 2016]. Disponível na Internet: <URL:http://hdl.handle.net/1822/17049>.

KAASE, Max (2013) - Research infrastructures in the social sciences: The long and winding road. In Kleiner, Brian [et al.] - *Understanding Research Infrastructures in the Social Sciences*. 1ª ed. Zurique : Seismo. ISBN 978-3-03777-133-4. Pt. 1, p. 19-27.

LAVOIE, Brian (2014) - *Meeting the challenges of digital preservation: The OAIS reference model*. [Em linha]. OCLC. [Consult. Out. 2016]. Disponível na Internet: <http://www.oclc.org/research/publications/library/2000/lavoie-oais.html>.

RCAAP. Sobre o RCAAP. *Missão e Objetivos*. [Em linha] Lisboa: FCT-FCCN. [Consult. Set. 2016] Disponível na Internet: <URL:http://projeto.rcaap.pt/index.php/lang-pt/sobre-o-rcaap/missao-e-objectivos>.

RENSCHLER, Isabelle; KLEINER, Brian; WERNLI, Boris (2013) - Concepts and key features for understanding social science research infrastructure. In Kleiner, Brian [et al.] - *Understanding Research Infrastructures in the Social Sciences*. 1ª ed. Zurique : Seismo. ISBN 978-3-03777-133-4. Pt. 1, p. 11-18.

RODRIGUES, Eloy [et al.] (2010) – *Os Repositórios de Dados Científicos: Estado da Arte*. [Em linha]. [Consult. Set. 2016]. Disponível na Internet: <URL:http://hdl.handle.net/1822/10830>.

VAN DEN EYNDEN, Veerle; BISHOP, Libby (2014) – *Incentives and motivations for sharing research data: researcher's perspectives* [Em linha]. 1ª ed. Essex : UK Data Archive, University of Essex. [Consult. 14 Jun. 2016]. Disponível na Internet: <URL:http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf>.

WACKEROW, Joachim; VARDIGAN, Mary (2013) – An established international metadata standard: The Data Documentation Initiative (DDI). In Kleiner, Brian [et al.] – *Understanding Research Infrastructures in the Social Sciences*. 1ª ed. Zurique : Seismo. ISBN 978-3-03777-133-4. Pt. 2, p. 158-167.

[i] A gestão de dados é aqui entendida como o conjunto de "práticas, manipulações, melhorias e processos que asseguram que os dados de investigação são de qualidade e que são bem organizados, documentados, preservados, sustentáveis, acessíveis e reutilizáveis" (Corti et al., 2014, p. 2).

[ii] De acordo com a *Digital Preservation Coalition*, entende-se por preservação digital o conjunto de atividades necessárias para garantir o acesso continuado à informação digital, mesmo em caso de falhas ou mudanças tecnológicas.